LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-PROC-215338

Growth Curve Models for the Analysis of Phenotype Arrays for a Systems Biology Overview of Yersinia pestis

I. K. Fodor, A. E. Holtz-Morris, S. L. McCutchen-Maloney

September 14, 2005

Joint Statistical Meetings
Minneapolis, MN, United States
August 6, 2005 through August 10, 2005

Growth curve models for the analysis of phenotype arrays for a systems biology overview of *Yersinia pestis*

IK Fodor, AE Holtz-Morris, and SL McCutchen-Maloney
Lawrence Livermore National Laboratory
P.O. Box 808, L-448, Livermore, CA 94551
fodor1@llnl.gov

September 8, 2005

Abstract

The Phenotype MicroArray technology of Biolog, Inc. (Hayward, CA) measures the respiration of cells as a function of time in thousands of microwells simultaneously, and thus provides a high-throughput means of studying cellular phenotypes. The microwells contain compounds involved in a number of biochemical pathways, as well as chemicals that test the sensitivity of cells against antibiotics and stress. While the PM experimental workflow is completely automated, statistical methods to analyze and interpret the data are lagging behind. To take full advantage of the technology, it is essential to develop efficient analytical methods to quantify the information in the complex datasets resulting from PM experiments. We propose the use of statistical growth-curve models to rigorously quantify observed differences in PM experiments, in the context of the growth and metabolism of *Yersinia pestis* cells grown under different physiological conditions. The information from PM experiments complement genomic and proteomic results and can be used to identify gene function and in drug development. Successful coupling of phenomics results with genomics and proteomics will lead to an unprecedented ability to characterize bacterial function at a systems biology level.

1 Phenotype microarrays

Bacterial phenotypes represent the cumulative function of biochemical pathways. Previous means to test bacterial phenotypes involved growing bacteria on specialized culture media on petri plates or in liquid broth cultures followed by visual inspection. Consequently, previous phenotype studies could realistically examine the effects of a few dozen chemicals at a time. The Phenotype MicroArray (PM) technology (Biolog Inc, Hayward, CA) revolutionized the field by introducing a high-throughput method that can simultaneously test thousands of phenotypes in an automated way [5, 7, 6, 23]. The PM system consists of two-dimensional micro plates, each of which contains 96 micro wells preloaded with separate chemicals. Aliquots of the microbe are incubated in the wells, and respiration is monitored automatically over time via the OmniLog robot. If the organism can metabolize a chemical, cells grow, indicating that the metabolic pathway for that chemical is functional. The resulting kinetic (growth curve) profiles are stored for subsequent analysis. Differences in the growth curves observed under varying physiological conditions, or from the growth of different organisms, indicate biological differences.

The current PM platform for microbial cells contains 20 micro plates, providing close to 2000 phenotype tests designed to cover known metabolic pathways. Eight micro plates contain compounds related to the main catabolic pathways for carbon, nitrogen, phosphorous and sulphur, as well as biosynthetic pathways. One plate tests osmotic stress factors and ion effects. Another investigates pH growth range and pH regulation. Ten plates test the sensitivity of cells to a number of chemicals such as antibiotics, anti-metabolites, membrane-active agents,

respiratory inhibitors, and toxic metals.

The OmniLog software included with the PM system allows rudimentary comparison of kinetic plots [4]. For each kinetic curve in an experiment, eight simple summaries such as the area under the curve, the maximum value, and the slope are calculated based on the raw data points. Three additional summaries are also provided based on the first-order derivatives.

Differences between growth curves in a well that correspond to bacteria grown under varying conditions, or to similar growth conditions of different strains, are evaluated by comparing a given summary statistics from the corresponding curves. If the difference between the summary statistics exceeds a user-selected threshold, a "significant" difference between the conditions or strains is declared. Replicate curves are averaged before the summary statistics are calculated. Comparisons are generally based on differences in the total area under the curves. Thresholds are manually selected, based on visual comparison of the data from all wells. While this methodology represents an unprecedented global means to study metabolic phenotypes, this analysis approach is time consuming and lacks statistical rigor.

To our knowledge, all PM results published to date relied on the threshold-driven method described in the previous paragraph to assess differences in the cells under investigation [13, 22, 21, 23, 2, 11, 14, 15]. In all cases, great emphasis was placed on the materials and methods, and the discussion sections. However, details about the analysis were scarce. Conclusions about phenotypic changes were based on "changes that were consistent and substantial" [7], on differences in the areas with undocumented thresholds [23], and on over 60% difference between summary statistics (presumable area, but this was not specified) [15]. In [22], the number of chemicals where differences were observed was given without an explanation of how they were determined.

Clearly, there is a gap between the advanced PM technology and our ability to make sense of the complex datasets emerging from it. To assess whether observed differences in the growth of cells on PM plates are statistically significant, a more rigorous approach is needed. We propose using statistical growth-curve models in Section 2, describe our preliminary results with the proposed methods in Section 3, and summarize in Section 4.

2 Proposed methods

Our solution seeks to improve the subjective methods currently used in the analysis of PM experiments. The approach is to develop an objective statistical growth-curve modeling framework [3, 20] and quantify automatically the significance of the results in a traditional hypothesis-testing context. The statistical methods will incorporate the information from the replicate curves and will attach a measure of significance to the results.

Data from published PM experiments, as well as our ongoing experiments, provide strong evidence that parametric models can be used to model the kinetic curves. Our initial results focus on the asymptotic regression (AR) and the four-parameter logistic (4PL) models, as these models fit most of the observations in our experiments. Figure 1 displays the two models.

The mathematical form of the AR model is given by

$$y^{AR}(x) = \Phi_1 + (\Phi_2 - \Phi_1)\exp[-\exp(\Phi_3)x], \qquad (1)$$

where $y^{AR}(x)$ is the growth at time x, Φ_1 is the asymptote (Asym), Φ_2 is the response at time 0 (resp0), and Φ_3 is the log rate constant (lrc), as illustrated in Figure 1(a). The time to reach half the maximum is the half-life, and is expressed as

$$t_{0.5} = \log(2)/\exp(\Phi_3). \qquad (2)$$

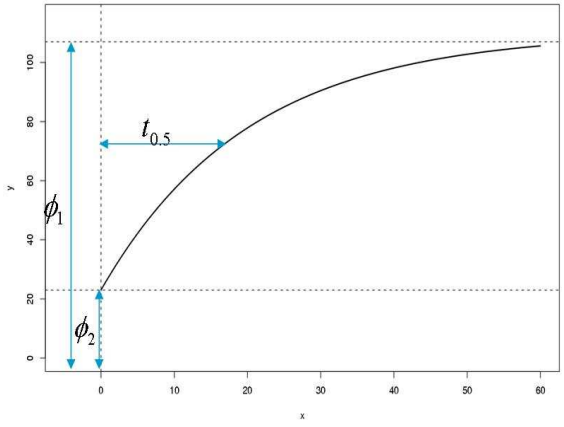
The 4PL model in Figure 1(b) is described by

$$y^{4PL}(x) = \Phi_2 + \frac{\Phi_1 - \Phi_2}{1 + \exp[(\Phi_3 - x)/\Phi_4]}, \qquad (3)$$

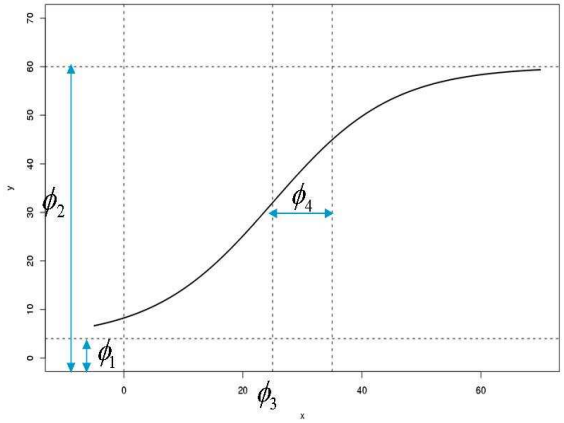
where where $y^{4PL}(x)$ is the growth at time x, Φ_1 is the lower asymptote (A), Φ_2 is the upper asymptote (B), Φ_3 is the midpoint between the asymptotes (xmid), and Φ_4 is a scale parameter such that when $x = \Phi_3 + \Phi_4$, the response is $\Phi_2 + (\Phi_1 - \Phi_2)/(1 + e^{-1})$ or about three-quarters of the distance from Φ_2 to Φ_1.

These parameterizations follow [18] and ensure the stability of the nonlinear fitting algorithms. Alternative parameterizations are also possible.

Given data from one well and a specific model, we estimate the corresponding model parameters and their uncertainties using nonlinear optimization algorithms [20, 19, 18]. Next, we construct confidence intervals for the differences in the parameter estimates that correspond to the



(a)



(b)

Figure 1: The (a) asymptotic regression and (b) four-parameter logistic models.

different conditions, and use hypothesis testing to assess whether the observed differences are statistically significant, at the desired significance level specified by the user.

3 Preliminary results

Yersinia pestis, the causative agent of plague, is a highly communicable bacterium that has been responsible for three historic pandemics with high mortality rates [17, 1]. In a recent pneumonic plague outbreak in the Republic of Congo, 57 of 130 patients had died as of March 30, 2005.

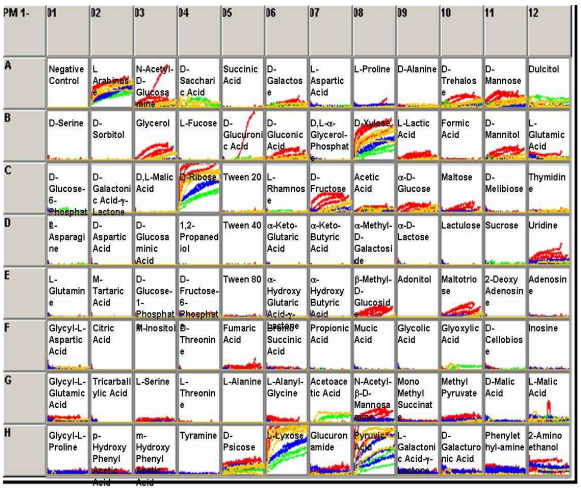


Figure 2: Phenotype of *Y. pestis* KIM5 D27 on the PM1 carbon source plate. Kinetic curves are shown in 96 wells under four physiologically relevant conditions: conditions mimicking the flea midgut (green), the mammalian host plasma (yellow), the mammalian intracellular (red), and an environmental control (blue). Three replicates per condition are plotted. The x-axis measures time in hours (288 time points for each replicate curve corresponding to measurements taken every fifteen minutes for three days), while the y-axis quantifies the growth in units of OmniLog signal intensity. The negative control values shown in the first well have been subtracted from each replicate.

Y. pestis virulence factors and a Type III secretion delivery system are induced thermally, when the bacterium en-

ters the mammalian host from the flea vector, and through host cell contact [9, 10]. To understand how growth conditions influence the virulence mechanism of the bacteria, we conducted a high-throughput experiment with PMs. *Y. pestis* was grown on 20 micro plates. We tested the cells under four biologically relevant growth conditions related to the flea vector and the mammalian host, and measured growth every fifteen minutes for three days. The calcium concentration and the temperature used represent the flea midgut (26°C, 4 mM Ca^{2+}), mammalian host plasma (37°C, 4 mM Ca^{2+}), and host intracellular (37°C, 0 mM Ca^{2+}) physiological conditions. An additional environmental control (26°C, 0 mM Ca^{2+}), was also tested. Three replicates were run under each of the four conditions. Figure 2 displays the data from the PM1 carbon source plate.

Figure 3 displays the results of fitting the two proposed models to two wells from the experiment. The first row illustrates an example with the AR model in Eq. (1). The replicate curves are shown with the color that corresponds to growth conditions, as shown in Figure 2. The black lines indicate the model fits for the four growth conditions, based on information from all the replicate curves. The second image in the first row presents the three parameter estimates (Asym, resp0, and lrc from Eq. (1) in the three columns, respectively) for the AR models fitted to the data collected under the four physiological growth conditions (shown in green, blue, yellow, and red). The intervals are the estimated 95% confidence intervals for the corresponding parameters. The results for the log rate constant (lrc) in the third column indicate that the cells grow significantly faster under the mammalian host intracellular (red) and mammalian plasma (yellow) conditions than under either the flea midgut (green) or control (blue) conditions. The corresponding half-life estimates of Eq. (2) are: 3.68 (red), 3.61 (yellow), 11.89 (green), and 19.36 (blue) hours.

The second row of Figure 3 presents an example result with the 4PL model in Eq. (3). The model is not adequate to describe the growth under the mammalian intracellular (red) condition, illustrating one of the challenges in automating the analysis. The second column presents the 95% confidence intervals for the four parameters in Eq. (3) for the three growth conditions where the model was appropriate. The flea midgut (green) and control (blue) conditions resulted in significantly larger asymptotes (B) and midpoints (xmid) than the mammalian plasma (yellow) condition.

4 Summary

We described the current state of the art in the analysis of phenotype microarrays, and demonstrated that more rigorous statistical methods can be used to realize the full potential of the technology. We suggested that statistical growth-curve modeling set in a hypothesis testing framework has the potential to add rigor to the interpretation of the phenotype array experiments. We presented preliminary results from a microbial experiment and demonstrated the applicability of the proposed methods.

Future plans include incorporating alternative models, model selection, improved hypothesis testing based on differences in the parameter estimates, as well as developing novel clustering algorithms to group the wells based on their growth curve profiles under the conditions of interest. Chemicals that cluster together are likely related or address a specific biochemical pathway, which will be helpful in the biological interpretation of the data. In addition to investigating traditional clustering methods based on the entire growth curves, we propose to develop more efficient algorithms based on the estimated growth curve models and parameter estimates.

Implementation of the proposed framework to the automated analysis of arbitrary phenotype array experiments will result in improved analysis of high-throughput phenotype experiments. This is the first step in an ambitious vision for a future bioinformatics platform for integrating results from from cross-technology applications. Ideally, improved results from phenotypic profiling experiments will be combined with findings using the more mature gene expression [16] and protein expression platforms [8, 12] in order to gain a better understanding of an organism at a systems biology level. In light of the imperfect technologies and unavoidable uncertainties, statistical methods are vital for a rigorous analysis of the data. Once proper statistical methods are in place to quantify the results of all the technologies involved, the information fusion can proceed. In turn, the combined results will facilitate greater mechanistic understanding of microbes, and contribute to the rapid detection and therapeutic intervention of infectious and emerging diseases.

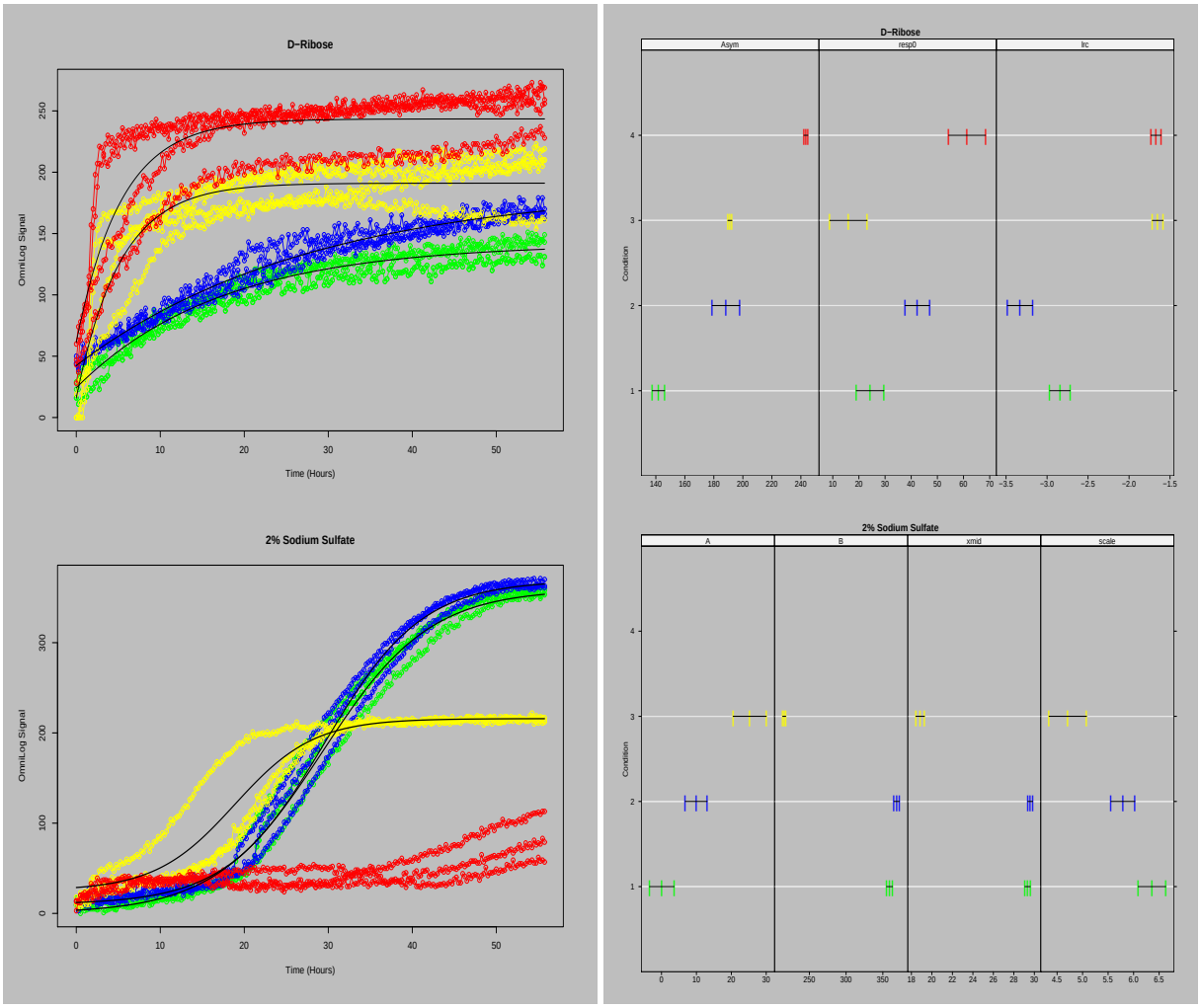


Figure 3: Selected results with the two parametric models. Top row: D-Ribose from the C04 position of the PM01 carbon sources plate. Bottom row: 2% Sodium Sulfate from the D05 position of the PM09 osmotic stress factors plate. First column: the replicates over the four physiological conditions (in color) and the model estimates (in black). The D-Ribose data was fit with the asymptotic regression model, while the Sodium Sulfate data with fit with the four-parameter logistic model. Second column: the 95% confidence intervals for the corresponding model parameters.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, with support from the Department of Homeland Security (Biological Countermeasures Program). This work was supported by Laboratory Directed Research and Development funding. UCRL-JC-.

References

[1] M. Achtman, G. Morelli, P. Zhu, T. Wirth, I. Diehl, B. Kusecek, A. J. Vogler, D. M. Wagner, C. J. Allender, W. R. Easterday, V. Chenal-Francisque, P. Wohrsham, N. R. Thomson, J. Parkhill, L. E. Lindler, E. Carniel, and P. Keim. Microevolution and history of the plague bacillus, Y*ersinia pestis*. *Proc Natl Acad Sci USA*, 101(51):17837–42, 2004.

[2] S. F. Al-Khaldi and M. M. Mossoba. Gene and bacterial identification using high-throughput technologies: genomics, proteomics, and phenomics. *Nutrition*, 20:32–8, 2004.

[3] D. M. Bates and D. G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, 1988.

[4] Biolog Inc. (Hayward, CA). *OmniLog Phenotype Microarray Software Release 1.2 User Guide*, 2003.

[5] B. R. Bochner. Sleuthing out bacterial identities. *Nature*, 339:157–8, 1989.

[6] B. R. Bochner. New technologies to assess genotype-phenotype relationships. *Nature Review Genetics*, 4:309–14, 2003.

[7] B. R. Bochner, P. Gadzinksi, and Panomitros E. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res*, 11:1246–55, 2001.

[8] B. A. Chromy, J. Perkins, J. L. Heidbrink, A. D. Gonzales, G. A. Murphy, P. A. Fitch, and S. L. McCutchen-Maloney. Proteomic characterization of host response to Y*ersinia pestis* and near neighbors. *Biochem Biophys Res Commun*, 320:474–9, 2004.

[9] G. R. Cornelis. Molecular and cell biology aspects of plague. *Proc Natl Acad Sci USA*, 97:8778–83, 2000.

[10] G. R. Cornelis. *Yersinia* type III secretion: send in the effectors. *J Cell Bio*, 158:401–8, 2002.

[11] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Pallson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429:92–6, 2004.

[12] C. Forde, J. M. Rocco, J. P. Fitch, and S. L. McCutchen-Maloney. Real-time characterization of virulence factor expression in Y*ersinia pestis* using a gfp reporter system. *Biochem Biophys Res Commun*, 324:795–800, 2004.

[13] L. Hamer et al. Gene discovery and gene function assignment in filamentous fungi. *Proc Natl Acad Sci USA*, 98(9):5110–5, 2001.

[14] C. G. Knight, R. Kassen, H. Hebestreit, and P. Rainey. Global analysis of predicted proteomes: functional adaptation of physical properties. *Proc Natl Acad Sci USA*, 101(22):8390–5, 2004.

[15] B. M. Koo, M. J. Yoon, C. R. Lee, T. W. Nam, Y. J. Choe, H. Jaffe, A. Peterskofsky, and Y. J. Seok. A novel fermentation/respiration switch protein regulated by enzyme IIA^{Glc} in E*scherichia coli*. *J Biological Chemistry*, 279(30):31613–21, 2004.

[16] V. L. Motin, A. M. Georgescu, J. P. Fitch, P. P. Gu, D. O. Nelson, S. L. Mabery, J. B. Garnham, B. A. Sokhansanj, L. L. Ott, M. A. Coleman, J. M. Elliott, L. M. Kegelmeyer, A. J. Wyrobeck, T. R. Slezak, R. R. Brubaker, and E. Garcia. Temporal global changes in gene expression during temperature transition in Y*ersinia pestis*. *J Bacteriol*, 186:6298–305, 2004.

[17] R. D. Perry and J. D. Fetherston. *Yersinia pestis* – etiological agent of plague. *Clin Microbiol Rev*, 10:35–66, 1997.

[18] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer: Statistics and Computing, New York, 2000.

[19] G. J. S. Ross. *Nonlinear Estimation*. Springer, 1990.

[20] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. Wiley, 1989.

[21] M. M. Tanzer, H. N. Arst, A. R. Skalchunes, M. Coffin, B. A. Darveaux, R. W. Heiniger, and J. R. Shuster. Global nutritional profiling for mutant and chemical mode-of-action analysis in filamentous fungi. *Funct Integr Genomics*, 3:160–70, 2003.

[22] B. S. Tracy, K. K. Edwards, and A. Eisenstark. Carbon and nitrogen substrate utilization by archival S*almonella typhimurium* LT2 cells. *BioMed Central Evolutionary Biology*, pages 2–14, September 8 2002.

[23] L. Zhou, X. H. Lei, B. R. Bochner, and B. L. Wanner. Phenotype microarray analysis of E*scherichia coli* K-12 mutants with deletions of all two-component systems. *J Bacteriology*, 185:4956–72, 2003.